0.1 Conditional preference-based energy learning with HTE embeddings and an LP economic oracle

Scenario encoding with a Hierarchical Temporal Encoder (HTE). Each scenario x (multilayer grid graph, exogenous drivers, and time series) is first encoded by a *Hierarchical Temporal Encoder* (HTE) into a context representation

$$h = \mathrm{HTE}_\phi(x), \qquad h \in \mathbb{R}^d, \tag{1}$$

where ϕ denotes the HTE parameters (trained beforehand or jointly, depending on the experimental setting). In practice, h may include multi-scale components (e.g., zone- and system-level embeddings) and temporal summaries; throughout this section we use h as a generic shorthand for the HTE-provided conditioning signal.

Decision space. The operational decision is represented by a high-dimensional binary vector $u \in \{0,1\}^M$ collecting unit-commitment (UC) flags, discrete demand-response (DR) activations, and discrete storage mode indicators across space and time. For sampling, we often operate in a continuous relaxation $\tilde{u} \in (0,1)^M$ (or in logit space) and binarize only before the physics-based refinement stage.

Conditional energy-based model (EBM). We learn a *conditional* energy function

$$E_\theta(u \mid h) \in \mathbb{R}, \tag{2}$$

where θ are the EBM parameters and h is the HTE embedding of the scenario. The EBM defines an implicit conditional distribution over discrete decisions:

$$p_\theta(u \mid h) \propto \exp(-E_\theta(u \mid h)). \tag{3}$$

Low-energy configurations are intended to correspond to low-cost operational strategies for scenario context h.

Implicit policy via normalized Langevin sampling. Given h, we generate a set of K candidate configurations by running a stochastic sampler targeting (3). Let $\mathcal{S}_\theta$ denote a normalized Langevin procedure operating in a relaxed space:

$$\{\tilde{u}^{(k)}\}_{k=1}^K \sim \mathcal{S}_\theta(h), \qquad \tilde{u}^{(k)} \in (0,1)^M, \tag{4}$$

and convert relaxed samples into binary decisions

$$u^{(k)} = \mathrm{Bin}(\tilde{u}^{(k)}) \in \{0,1\}^M, \tag{5}$$

where $\mathrm{Bin}(\cdot)$ is thresholding or Bernoulli sampling.

Hierarchical feasibility decoder and LP worker (economic oracle). Each discrete candidate $u^{(k)}$ is passed through a hierarchical decoder $\mathcal{D}$ that produces a coherent continuous initialization (dispatch, storage trajectories, DR profiles, and flows) consistent with $u^{(k)}$:

$$y_{\mathrm{dec}}^{(k)} = \mathcal{D}\Big(x, u^{(k)}\Big). \tag{6}$$

Then, an LP worker $\mathcal{W}$ *hard-fixes* the discrete decisions $u^{(k)}$ and solves a continuous dispatch problem to validate feasibility and compute the realized operational cost:

$$(\mathrm{feas}^{(k)},\, y_\star^{(k)},\, C^{(k)}) = \mathcal{W}\Big(x, u^{(k)}, y_{\mathrm{dec}}^{(k)}\Big), \tag{7}$$

where $\mathrm{feas}^{(k)} \in \{0,1\}$ denotes feasibility and $C^{(k)} \in \mathbb{R}_+$ is the resulting cost (including penalty terms such as VOLL when applicable). Crucially, $\mathcal{W}$ is treated as a *non-differentiable* physics-aware oracle: it provides exact economic feedback, but gradients are *not* propagated through the LP.

Preference signal induced by realized costs. Among feasible candidates we select the lowest-cost solution

$$k^\star \in \arg\min_{k:\,\mathrm{feas}^{(k)}=1} C^{(k)}, \qquad \hat{u} := u^{(k^\star)}. \tag{8}$$

When available, we also obtain a reference solution u^+ from the MILP oracle (optimal or best incumbent under a time limit) with cost C^+. The LP worker thus induces an ordering over decisions:

$$u^{(i)} \succ u^{(j)} \quad \Longleftrightarrow \quad C^{(i)} < C^{(j)}. \tag{9}$$

Learning then aims at shaping the energy landscape so that lower-cost decisions receive lower energy under the same conditioning h.

Preference-based objective (conditional energy shaping). We train the conditional energy $E_\theta(\cdot \mid h)$ using a margin-ranking loss comparing the MILP reference u^+ to hard candidates produced by the pipeline. Let $\mathcal{K}$ be a set of "hard negatives" (e.g., feasible candidates with high realized costs, or a diverse subset). We minimize

$$\mathcal{L}_{\mathrm{rank}}(\theta) = \frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} \max\Big(0,\; m + E_\theta(u^+ \mid h) - E_\theta(u^{(k)} \mid h)\Big), \tag{10}$$

where $m > 0$ is a margin. To emphasize rare but catastrophic failures, we optionally use a cost-aware weighting based on the realized cost gap:

$$w_k = \mathrm{clip}\Big(\log\big(1 + (C^{(k)} - C^+)_+\big),\; 0,\; w_{\max}\Big), \tag{11}$$

and define

$$\mathcal{L}_{\text{w-rank}}(\theta) = \frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} (1 + \alpha w_k)\, \max\Big(0,\; m + E_\theta(u^+ \mid h) - E_\theta(u^{(k)} \mid h)\Big), \tag{12}$$

with $\alpha \geq 0$.

Gradient flow and training protocol. The full pipeline (Langevin sampler $\rightarrow$ decoder $\rightarrow$ LP worker) is used to generate economically meaningful hard candidates and costs. Gradients are computed *only* through the EBM evaluations $E_\theta(\cdot \mid h)$ in Eqs. (10)–(12). In particular, we do not differentiate through the decoder $\mathcal{D}$, the LP worker $\mathcal{W}$, nor the MILP oracle. The conditioning by HTE embeddings $h = \mathrm{HTE}_\phi(x)$ ensures that the learned energy landscape adapts to scenario-specific spatio-temporal patterns (demand, VRE, congestion, storage tension), enabling fast inference via sampling while retaining physics-aware economic evaluation through the LP worker.

Algorithm 1: Conditional preference-based EBM training (HTE-conditioned) with an LP economic oracle

Input: Scenario dataset $\{x_n\}$; HTE encoder HTE_ϕ; EBM E_θ
Input: Sampler $\mathcal{S}_\theta$; decoder $\mathcal{D}$; LP worker $\mathcal{W}$
Input: Hyperparameters: K candidates, margin m, weighting α

for *each minibatch of scenarios* x **do**
  Compute conditioning embeddings $h \leftarrow \mathrm{HTE}_\phi(x)$;
  Sample $\{\tilde{u}^{(k)}\}_{k=1}^{K} \sim \mathcal{S}_\theta(h)$;
  Binarize $u^{(k)} \leftarrow \mathrm{Bin}(\tilde{u}^{(k)})$;
  for $k = 1$ **to** K **do**
    $y_{\mathrm{dec}}^{(k)} \leftarrow \mathcal{D}(x, u^{(k)})$;
    $(\mathrm{feas}^{(k)}, C^{(k)}) \leftarrow \mathcal{W}(x, u^{(k)}, y_{\mathrm{dec}}^{(k)})$;
  Select a hard-negative set $\mathcal{K}$ from feasible candidates (e.g., high-cost feasible);
  Retrieve MILP reference (u^+, C^+) for these scenarios when available;
  Compute $\mathcal{L}_{\text{w-rank}}$ using energies $E_\theta(\cdot \mid h)$ only;
  Update θ by backpropagation through E_θ (no gradients through $\mathcal{D}$ or $\mathcal{W}$);